2.1 Data Collection Techniques

At times, you may want to use information collected in one system or database in other formats. This may be done to share data between locations, utilize another software package for specialized data manipulation, or export data for use in reports or other documents.

Generally, to the extent possible, the creation of duplicate databases is discouraged. If the source data are changing, the destination data will need to be updated periodically to ensure that current data are being accessed. However, on other occasions, it is useful to take a "snapshot" of the data by copying the desired records and fields into a separate file. This allows the user to manipulate the data to obtain counts and trends without the data changing between operations.

The following sections will discuss items that should be considered in sharing data between systems or software packages. The first section will discuss general considerations in downloading data. The second section deals specifically with downloading data from the DOE Performance Indicator Data System (PIDS).

Guide for Downloading Data

There are two different approaches to transferring data between systems. The first consists of transferring formatted files, e.g., a .DBF file, while the second consists of transferring a formatted ASCII text file that can be imported into the destination system. The preferred method depends on the capability of the host and destination databases and the desired information to be transferred.

Many database and spreadsheet software packages will work with a number of different file formats. In some cases, the file formats are directly compatible between software packages. For example, FoxPro will directly load or accept a .DBF file created with dBASE. In other cases, a conversion utility is required. For example, the Lotus 123 translate utility will convert a number of different database and spreadsheet formats into a file that can be loaded directly into Lotus 123. In many cases, these translation utilities are not directly available from the main application menu but must be run from a higher level menu or directly from the DOS prompt. The most common formats that are directly compatible with multiple software packages are the WKS format for spreadsheet programs and the DBF format for databases.

If a format is not available that can be moved directly between software packages, or translated to work with another package, most software packages will export and import a formatted ASCII data file. The two most commonly used types of formatted ASCII data files used in transferring data between applications are the standard data format (SDF) and the character delimited format (CDF).

In the SDF format, each row contains one record and each field is a predefined size. No punctuation is used in the records and each record ends with a carriage return and line feed. This format is particularly useful for working with columnar data. An example of a file in SDF format is shown below.

```
Smith      Tom      25      234-43-5547

Jones      John     41      442-78-4531
```

The CDF format is widely used in spreadsheets and databases. In this format, each row again contains one record. However, fields are separated by a character, usually a comma, and character data are enclosed by punctuation marks, usually a double quote. Leading or trailing blanks in the data are trimmed off (i.e., fields may be varying lengths). Again, each record ends with a carriage return and line feed.

The above file in CDF format, with comma separators and double quote delimiters, would be captured as:

```
"Smith", "Tom", 25, "234-43-5547"

"Jones", "John", 41, "442-78-4531"
```

Although commas are most commonly used as separators and double quotes as delimiters, many software packages allow the user to specify the characters that are used. In some cases, use of an alternate character may be preferable. For example, FoxPro will output a tab delimited file that is particularly useful for outputting data that will be imported into a word processor.

It should be noted that, in some software packages, the file extension is important when creating a text file for importing into a software package. For example, dBASE requires a .TXT extension for files that are being imported.

Some general hints on downloading data from one system:

1. Make a backup copy of the downloaded file and the file it will be imported into BEFORE you do anything else.
2. Many database systems include some type of a unique identifying field for each record, e.g., an index number. If this field is accessible, including it in the download may be useful to facilitate later downloads to update information or include additional fields from the host database.

3. When data are moved between software packages using formatted ASCII files, each field will be imported into a different column in a spreadsheet or into a different field in a database. If you are importing data into an existing database or spreadsheet, the order of the fields in the downloaded file needs to match the order of the fields in the existing database. The field sizes in the existing database must be at least as large as the corresponding fields in the downloaded data.
4. "Layered" spreadsheets (e.g., multidimensional .WK3 worksheets) do not import well into other programs. Frequently the layers are lost. It is better to save the spreadsheet in a different format and then try to import it.
5. Print a sample of the downloaded file (use a small font). Even though you can view the file on screen, some problems in data continuity are more apparent on the printed page.
6. If the file to be imported contains a large number of fields or extensive narrative data, keep in mind that each line in the file will be treated as a separate record. Depending on the system or the software package being used, there may be a limit to the line length. If this is the case, the data for a single record may wrap to more than one line, and subsequent lines need to be associated with the main record line during the import process.

Downloading Data from PIDs

The current design of DOE's Performance Indicator Data System (PIDS) provides the capability of downloading data in a formatted ASCII data file (CDF format). This data file is compatible with most spreadsheet and database programs.

The delimited ASCII file download option from the PIDS report option automatically creates a file in the same format as the upload file that is used for submitting data to PIDS. Each line is a separate record. Fields are separated by commas, and all data are delimited with double quotes. All data in PIDS are treated as character data. The format for a performance indicator (PI) data record in PIDS is as follows:

"year-quarter", "facility or contractor", "PI", "PI value", "change flag", "PI narrative"

The characteristics of the fields are as follows:

Year-Quarter	Character (4)
Facility	Character (20)
PI number	Character (8)
PI value	Character (14)
Change flag	Character (2)
Narrative	Character (Unlimited)

PI numbers or identifiers are stored without decimal points, e.g., PI 1.2 is stored as 12.

All PI values are stored in PIDS as character data. In some cases, data may not be available. In the records where values are not available, the value is replaced by a code as follows:

-1 Currently unavailable (CU)
-2 Not available - security concerns (NAS)
-3 Not applicable (NA)

The format for a PIDS root cause (RC) data record is as follows:

"year-quarter", "facility or contractor", "RPI", "root cause", "RC value", "change flat", "PI narrative".

The characteristics of the fields are as follows:

Year-Quarter	Character (4)
Facility	Character (20)
RPI number	Character (8)
Root cause	Character (4)
RC value	Character (14)
Change flag	Character (2)
Narrative	Character (Unlimited)

The RPI number is the PI number or identifier, preceded by the letter "R", e.g. R12.
When an error is detected after the data submission deadline, an errata form must be approved and submitted in order for data to be changed in PIDS. The format for an errata record is as follows:

"year-quarter", "facility or contractor", "PI", "old PI value", "new PI value", "change flag", "PI narrative", "errata basis".

The characteristics of the fields are as follows:

Year-Quarter	Character (4)
Facility	Character (20)
PI number	Character (8)
Old PI Value	Character (14)
New PI value	Character (14)
Change flag	Character (2)
Narrative	Character (Unlimited)
Errata basis	Character (Unlimited)

Use of the Internet

The use of the Internet has also become a valuable tool to collect, capture, and share information from different sources. Internet provides many capabilities, including the capability to transfer data files electronically. Large amounts of data can be transferred from one location to another in a matter of seconds. This capability can improve the timeliness of obtaining information necessary to support organizational performance measurement analyses. Many books and manuals are available that provide information on use of Internet.